EmoChildRu: Emotional Child Russian Speech Corpus

Elena Lyakso[1], Olga Frolova[1], Evgeniya Dmitrieva[1], Aleksey Grigorev[1], Heysem Kaya[2], Albert Ali Salah[2], and Alexey Karpov[3,4]

[1] The Child Speech Research Group, St. Petersburg State University, Russia
[2] Department of Computer Engineering, Bogazici University, Istanbul, Turkey
[3] St. Petersburg Institute for Informatics and Automation of RAS, Russia
[4] ITMO University, St. Petersburg, Russia
lyakso@gmail.com,heysem@boun.edu.tr,karpov@iias.spb.su

Abstract. We present the first child emotional speech corpus in Russian, called "EmoChildRu", which contains audio materials of 3-7 year old kids. The database includes over 20K recordings (approx. 30 hours), collected from 100 children. Recordings were carried out in three controlled settings by creating different emotional states for children: playing with a standard set of toys; repetition of words from a toy-parrot in a game store setting; watching a cartoon and retelling of the story, respectively. This corpus is designed to study the reflection of the emotional state in the characteristics of voice and speech and for studies of the formation of emotional states in ontogenesis. A portion of the corpus is annotated for three emotional states (discomfort, neutral, comfort). Additional data include brain activity measurements (original EEG, evoked potentials records), the results of the adult listeners analysis of child speech, questionnaires, and description of dialogues. The paper reports two child emotional speech analysis experiments on the corpus: by adult listeners (humans) and by an automatic classifier (machine), respectively. Automatic classification results are very similar to human perception, although the accuracy is below 55% for both, showing the difficulty of child emotion recognition from speech under naturalistic conditions.

Keywords: emotional child speech, perceptual analysis, spectrographic analysis, emotional states, computational paralinguistics.

1 Introduction

Speech databases are an indispensable part of speech research. Their structure and technical characteristics depend on specific tasks and on the aims of the investigation for which the data are collected. One of the important areas of speech studies is the detection of the speakers emotional state in voice and speech. Emotions play an important role in communication, being one of the major factors of human behavior and indicative of a person's mental states.

For the study of emotional speech, it is necessary to use special corpora. There are very few child emotional speech databases available for child speech

research community. For example, such corpora exist for English, German, Italian, and Swedish [1] pre-school and school children. There is also a child speech database containing a large vocabulary rated on emotional valence (positive, neutral, and negative) by French children, differing in both age (5-9 years old) and sex (girls and boys) [13]. However, there are no databases containing emotional child speech material produced by Russian children.

For the Russian language studies, we created the corpus called "INFANTRU", which is the first database containing vocalizations of infants and children for a Slavic language [7]. The database contains 2967 recordings (70 hours) of 99 children aged between 3 to 36 months. For 76 children, the corpus contains the longitudinal vocalization data from 3 to 36 months. A second corpus, called "CHILDRU", stores speech material for 4 to 7 year old Russian children. This database holds 28079 recordings (20 hours in total) of 150 Russian children: 142 children growing in families and 8 orphans, growing in a child care facility [7].

Creation of a corpus of child emotional speech is more difficult than the construc-tion of corpora of emotional speech of adults. In the case of adults, actors are often involved to portray the necessary emotional conditions [5], or records of patients from a psychiatry clinic are used. Such approaches cannot be used for children. It is necessary to model communicative tasks in which the child is not conscious of being recorded to produce veridical emotional reactions. The creation of the corpus should be based on a verified and clear method of obtaining spontaneous speech manifestations of certain emotional reactions. By nature, collection of child emotional speech data is 'in the wild'.

It is well known that acoustic and linguistic characteristics of child speech are widely different from those of adult speech. The child speech is characterized by a higher pitch value, formant frequencies and specific indistinct articulation with respect to the adult speech. It was recently shown that adult Russians could recognize 60% to 100% of 4-5 years old childrens words and phrases during calm, spontaneous speech, and the amount of words recognized by adult native speakers does not increase much for children's speech from 4-5 to 7 years [4]. It is obvious that analysis of child speech must be separately investigated. The contributions of our current work are: 1) the collection of emotional speech material for 3-7 year old children in the form of a corpus and 2) a preliminary analysis of these data for studies of emotion development.

2 Emotional Child Russian Speech Corpus - EmoChildRu

2.1 Data collection

"EmoChildRu" is the first database containing emotional speech material from 3-7 year old Russian children. The database includes 20.340 recordings (30 hours) of 100 children growing in families. All children were born and live in the city of St. Petersburg (parents of the children were also born in St. Petersburg, or have been living there for more than 10 years). Places of recording were at home, in laboratory and kindergarten. The three different recording conditions are playing

with a standard set of toys, repetition of words from a toy-parrot in a game store setting, and watching a Russian cartoon called "Masha and bear" from iPad and the retelling of the story, respectively. All experiments had a duration of 2 minutes. Every record is accompanied by a detailed protocol and video recording of childs behavior in parallel. The speech materials are grouped based on the protocol and the recording situation, in accordance with underlying base emotions: sadness, anger, fear, gladness. So far, about 10% of the data are annotated for emotional states. The database contains additional information about the child's psychophysiological condition before and after speech recording. Original EEG and evoked potential (EP) records (visual stimuli images of the facial expression of infants and 3-4 year old children), dialogue descriptions, speech developmental and cognitive scale data are included in the database whenever available. The recordings were made with a Marantz PMD660 digital recorder and with a single "SENNHEISER e835S" external microphone. The speech sounds were analyzed by experts, using Syntrillium's "Cool Edit Pro" sound editor. Stressed vowels were selected from all phrases stressed words. The pitch as well as the vowel duration and phrase prosody were calculated.

2.2 Corpus and software structure

The corpus and accompanying software package includes the database, as well as a shell component to navigate, browse and search information in the database. The database stores information about files (file type, size in bytes, duration in minutes and seconds, name, description, etc.) and their relationships. Speech files are in Windows PCM format, 22050 Hz, 16 bit per sample.

A software tool was developed for enabling the experts to work with the "EmoChildRu" corpus. The shell component is developed in Microsoft Visual C#, and it is designed to enable working with the database under Windows operating system. This software also allows choosing speech material using a query interface, along dimensions such as a type of emotional state, child age and gender. It allows choosing any or each available feature, including speech material and acoustic characteristics of speech in different emotional states, video elements of nonverbal behavior, evoked potentials data on visual "emotional stimuli", or ECG data; filtered for all children of a certain age or gender.

3 Data Analysis

Three test sequences were formed from the speech material of 30 children aged from 3 to 7 years. The test sequences were composed into three groups (3-4 year old, 5 year old, 6-7 year old children) and every test sequence includes 10 phrases uttered by children in a comfortable emotional state, 10 phrases in a discomfort state and 10 phrases in a neutral (calm) state. We have used 90 sequences in total for testing. The child's emotional state was revealed based on the recording setting and by analysis of the video fragments by five speech experts. The test sequences were presented to 100 adults (native Russian speakers) for perceptual

analysis. Thus, each phrase was evaluated by 100 auditors, and the correctness of auditors in correctly recognizing the emotional state on the base of speech samples (perception rate) was calculated. The aim of the perceptual analysis was to investigate how correctly the child's emotional state was perceived by native speakers. Spectrographic analysis of speech samples from the test sequences was also carried out.

4 Experimental Results

4.1 Human recognition of emotional states

At first, we studied recognition of a child's emotional state in speech material by humans. In average, adult native speakers correctly recognized less than 50% of speech material samples in the test sequences (Fig. 1A).

Both discomfort and comfort conditions in the speech of 3-5 year old children were recognized by adults with the perception rate of 0.75-1.0 better compared to the neutral condition. Humans' emotion recognition rate for the 6-7 year old children was higher than for the 4 year old children, as shown in Figure 1B ($p<0.01$, Mann-Whitney test).

Spectrographic analysis revealed that speech samples interpreted by humans as discomfort, neutral and comfort are characterized by specific sets of acoustic features (Fig. 2). Discomfort speech samples are characterized by phrase duration lengthening; highest pitch values (comparatively) in phrase and in stressed vowels selected from stressed words in the phrase ($p<0.05$ vs. neutral state, Mann-Whitney test); an increase of minimum pitch value in phrases; an increase of the pitch range in stressed vowels of stressed words in the phrase ($p<0.05$ vs. neutral state); falling pitch contour of stressed vowels of stressed words in the phrase. Comfort speech phrases have short duration, together with long stressed vowel duration; pitch values of phrases are increased, but less so

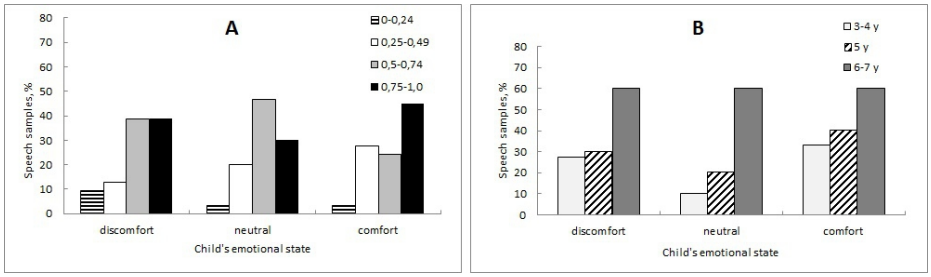

Fig. 1. Percentages of emotional child speech samples perceived by adult native speakers: A - correctly recognized as discomfort, neutral and comfort with the perception rate of 0-0.24 (horizontal hatch), with the rate of 0.25-0.49 (white color bar), with the rate of 0.5-0.74 (light gray) and with the rate of 0.75-1 (black); B - correctly recognized with the rate of 0.75-1.0 at different childrens ages: 3-4 years (light gray), 5 years (sloping hatch), 6-7 years old (grey).

compared to discomfort samples; pitch range in the stressed vowels is similar to discomfort samples; rising pitch contours of stressed vowels. Neutral speech samples are characterized by lowest values of vowel duration, stressed vowels' pitch and pitch range; and flat pitch contour is dominated as well (Fig. 2, Table 1).

Most speech samples that were correctly recognized by humans in the experiment have a complex shape of phrase pitch contours (>70% samples). The analysis of features of all stressed vowels from stressed words revealed that discomfort speech samples have mainly a falling shape, while comfort speech samples have a rising shape (Table 1).

Table 1. Distribution of pitch contour shapes for correctly recognized speech samples

Children's state	Age	Pitch contour shape, %				
		flat	rising	falling	U-shaped	bell-shaped
discomfort	3-4	0	33	**67**	0	0
	5	0	0	**100**	0	0
	6-7	33	0	**67**	0	0
neutral	3-4	100	0	0	0	0
	5	0	0	100	0	0
	6-7	67	0	16.5	0	16.5
comfort	3-4	0	**67**	0	33	0
	5	0	**75**	25	0	0
	6-7	17	50	0	33	0

Almost all neutral speech samples have flat, falling and bell-shaped pitch contours, and the first two patterns are the most common. U-shaped pitch contour is revealed in comfort speech samples only. Variety of pitch contour shapes in stressed vowels increases by 6-7 years, compared to younger children. It was revealed that the duration of phrases increases, and the duration of stressed vowels and pitch values decreases with increasing age of the children. The differences in the acoustic characteristics of speech samples correctly recognized as discomfort, neutral and comfort are more expressed at the age of 3-5 years. Correctly recognized speech samples of 6-7 years old children do not differ sig-

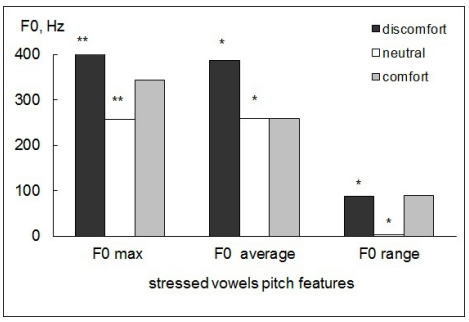

Fig. 2. The pitch values and the duration of vocalizations recognized by auditors as different emotional infants and chimpanzees states.

nificantly in acoustic features. Adult listeners mostly rely on the meaning of the phrase. Analysis of speech samples correctly recognized by adults revealed that detection of word meanings from these phrases improved with increase of child's age: meaning of words from 57% of child's phrases were detected at the age of 3-4 years, 86% - at the age of 5 years, 100% - at the age of 6-7 years.

4.2 Automatic classification of emotional states

Automatic processing of child speech is tackled in many recent studies, but automatic detection and classification of emotional states in children's speech in the wild is a new direction of research. As our perception analysis reveals, the recognition of children's emotions from speech is hard, even though some discriminative prosodic patterns can be discerned. The overall human recognition accuracy is about 50% for the three-class problem.

In the second experiment, we employ an objective, automatic classification algorithm. For this purpose, we use a subset of the corpus, where all speech files have 1 to 5 seconds of speech and all five child speech experts agree on the emotion annotation. Note that the annotation is done on the audio-visual material, including the linguistic information, although only acoustic features are used for automatic classification. The subset is collected from 50 children, and the number of speech files per child ranges from 1 to 78 (11.7 files on average, with a standard deviation of 12.5) increasing the difficulty of age/gender balanced partitioning, as well as recognition. There are 23 boys and 27 girls in the selected set, Meanstd age is 5.1 ± 1.1 years. Trying to keep the distribution of emotion labels as balanced as possible, we obtain the partitioning shown in Table 2. We report training and testing classification results in terms of accuracy and Unweighted Average Recall (UAR), which is introduced as performance measure in the INTERSPEECH 2009 Emotion challenge [12]. UAR is used to overcome the biased calculation of accuracy towards the majority class. It also gives a chance-level baseline performance as 1/K, where K is the number of classes. In our case the chance-level is 33.3%.

We extract openSMILE [2] features using a configuration file used in the INTERSPEECH (IS) ComParE Challenges in 2010 [10] and 2013 [11]. These feature sets contain 1582 and 6373 suprasegmental features, respectively. Utterance features are obtained by passing descriptive functionals (e.g. moments, percentiles, regression coefficients) on the Low Level Descriptors (e.g. pitch, MFCC 1-14, jitter, etc.). For classification, we use Linear SVM implementation from the WEKA data mining tool [3] and train models with the Sequential Minimal Optimization Algorithm [8]. To avoid over-fitting, we leave the SVM complex-

Table 2. Distribution of classes and gender (M/F) in train and test partitions

	M/F	Comfort	Neutral	Discomfort	Total
Train	16/20	144	164	52	**360**
Test	7/7	90	88	47	**225**
Total	**23/27**	**234**	**252**	**99**	**585**

Table 3. Classification results for 3 valence states of comfort, discomfort, and neutral

	IS 2010 features		IS 2013 features	
Preprocessing	**Accuracy**	**UAR**	**Accuracy**	**UAR**
Z-norm	**53.3%**	**55.0**	46.7%	45.3
Minmax-norm	52.4%	50.0	47.1%	45.7

ity parameter at its default value of 1. The classification can be thought of as a three-state valence classification problem. It is well known that valence classification from acoustics is poorer compared to arousal classification, and is at almost chance level in challenging conditions without adaptation (e.g. cross-corpus setting) [9]. In Table 3, we observe better classification performance with IS 2010 features compared to IS 2013 features, and the automatic approach performs better than human perception.

5 Discussion

The results of our previous study [6] showed the extent of the ability of adults in proper recognition of emotional states of infant vocalizations. Speech experts recognize discomfort vocalizations of infants better than vocalizations that reflect a comfortable condition. In the present study, we aimed to analyze the emotional speech of children 3-7 years of age. Choosing the age range as 3-7 years is due to the evolution of the grammatical skills of speech at 4 years and the ability of effective communication of a child with an adult, including regulation of emotional expressions, the "truth" of emotion and the contribution of society in the organization of the child's behavior is comparatively small. The upper bound age of 7 years is associated with the end of the preschool time of children in the Russian Federation. There is still no systematic training of kids at school, but more stable brain activity, compared to the earlier ages. There are only a few databases with emotional speech of children before 4 years of age [1].

The presented experimental results show that lexical information has more discriminative power in recognition of valence compared to acoustic features for speech samples of 6-7 year old children. Human perception of emotion is higher in the speech of older children, which has several implications. It is harder to recognize sentiment with younger children, since linguistic and acoustic control skills are not mature enough. Despite this issue, analyzing/monitoring child emotion in early ages is important not only for linguistics, but also for analysis of neurological development/disorders. The preliminary automatic classification study reveals very close performance to human perception, which can be taken as a gold standard. As in cross-corpus acoustic emotion recognition [9], we observed improved performance due to within-set normalization scheme.

6 Conclusions

In this paper, we introduced the EmoChildRu corpus that has been designed to study the reflection of the emotional state in the characteristics of voice and

speech and for studies of the formation of emotional states in ontogenesis. From our point of view, child emotional speech material in our database can be the basis for scientific projects on how the Russian language is mastered by children and on the emotional development of children. This corpus can also be used for the research and development of automated child speech recognition systems.

Acknowledgments. This study is financially supported by the Russian Foundation for Humanities (project # 13-06-00041a), the Russian Foundation for Basic Research (projects # 13-06-00281a, 15-06-07852a, and 15-07-04415a), the Council for grants of the President of Russia (project # MD-3035.2015.8) and by the Government of Russia (grant No. 074-U01).

References

1. Batliner, A., Blomberg, M., D'Arcy, S., Elenius, D., Giuliani, D., Gerosa, M., Hacker, C., Russell, M.J., Steidl, S., Wong, M.: The pf_star children's speech corpus. In: INTERSPEECH. pp. 2761–2764 (2005)
2. Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the international conference on Multimedia. pp. 1459–1462. ACM (2010)
3. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. ACM SIGKDD explorations newsletter 11(1), 10–18 (2009)
4. Lyakso, E., Frolova, O., Grigoriev, A.: Acoustic characteristics of vowels in 6 and 7 years old russian children. In: Proc. International Conference INTERSPEECH. pp. 1739–1742 (2009)
5. Lyakso, E.: Study reflects the voice of emotional states: Comparative analysis chimpanzee, human infants and adults. In: Proc. XVI European Conference on Development Psychology ECDP-2013 (2013)
6. Lyakso, E., Grigorev, A., Kurazova, A., Ogorodnikova, E.: "INFANT. MAVS"-multimedia model for infants cognitive and emotional development study. In: Speech and Computer, pp. 284–291. Springer (2014)
7. Lyakso, E.E., Frolova, O.V., Kurazhova, A.V., Gaikova, J.S.: Russian infants and children's sounds and speech corpuses for language acquisition studies. In: Proc. International Conference INTERSPEECH. pp. 1878–1881 (2010)
8. Platt, J., et al.: Fast training of support vector machines using sequential minimal optimization. Advances in kernel methods: support vector learning 3 (1999)
9. Schuller, B., et al.: Cross-corpus acoustic emotion recognition: variances and strategies. Affective Computing, IEEE Transactions on 1(2), 119–131 (2010)
10. Schuller, B., et al.: The interspeech 2010 paralinguistic challenge. In: INTERSPEECH. pp. 2794–2797 (2010)
11. Schuller, B., et al.: The interspeech 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism (2013)
12. Schuller, B., Steidl, S., Batliner, A.: The interspeech 2009 emotion challenge. In: INTERSPEECH. vol. 2009, pp. 312–315 (2009)
13. Syssau, A., Monnier, C.: Children's emotional norms for 600 french words. Behavior Research Methods 41(1), 213–219 (2009)